

Research on Digital Preservation
within projects co-funded by
the European Union in the ICT programme

Stephan Strodl, Vienna University of Technology, Austria

Petar Petrov, Vienna University of Technology, Austria

Andreas Rauber, Vienna University of Technology, Austria

Contributions:

Ross King (AIT, Austria)

Rainer Schmidt (AIT, Austria)

Christoph Becker (Vienna University of Technology, Austria)

Mark Guttenbrunner (Vienna University of Technology, Austria)

May 2011

Disclaimer

This report provides an overview about the research on Digital Preservation of initiatives co-funded by the European Commission in the ICT programme. A subject as complex as the topic research can be seen from many points of view. In order to open new perspectives, we tried to give structured summaries of the activities from different angles.

The range of research projects and the large number of deliverables, publications and outcomes make this report a challenging task. The representation of any single project cannot be complete. We tried to consider all major contributions of the research projects in the area of digital preservation. All material used for this report is publicly available.

The content of this document does not necessarily reflect the opinion of the European Commission.

Acknowledgment

The authors of this report would like to thank the following persons for their comments on this report. Their comments made the useful parts of this report better but any remaining misinterpretations or mistakes are those of the authors.

Vangelis Banos, Elisabeth Freyre, David Giaretta, Mariella Guer-
cio, Matthias Hemmje, Seamus Ross, Eloit Salant and Daniel
Teruggi

Executive Summary

This report gives an overview about research on digital preservation throughout projects sponsored by the European Commission as part of the sixth and seventh framework programmes for research and technological development. It summarises the objectives, developments, similarities and differences as well as accomplishments and results of these projects. The report also considers research agendas in the field of digital preservation and identifies challenges for the future.

The Past

The first efforts in digital preservation on a European level were focused on raising awareness for long term preservation issues. Starting with ERPANET, continued with DELOS and DigitalPreservationEurope, a series of presentations and workshops were held with the focus of awareness raising, identification of potential target groups and creation of a scientific community addressing collaboratively this novel and interdisciplinary topic. A number of fact sheets and briefing papers were published addressing different settings of digital preservation. The second aim was the consolidation of the existing work in the area of digital preservation, initially integrating national initiatives, later including the different research projects on a European level. One result of this work was the establishment of the WePreserve initiative.

The beginning of the scientific work is shaped by the establishment of common problem definitions, terminology and concepts. Influenced by the library and archive community, first models and tools were developed (for example DELOS DPC Testbed [13]). The work focused on topics supporting digital preservation such as metadata standards, system concepts, selection and appraisal policies and format identification. The research in digital preservation was primarily focused on office documents and images in institutional settings. The issue of preservation of non-traditional objects was recognised on a very early stage, but it was addressed in research projects at a later point in time. In a next phase, a series of research projects targeted more technical aspects and actual tool and framework development of digital preservation (Planets, Caspar, Shaman, DigitalPreservationEurope and Protage). This led to the availability of concrete solutions, as well as a solid body of expertise both on a theoretical as well as an applied level via a series of case studies. It has also shown its impact via the influence these results have on international standardization initiatives with strong European presence (e.g. PREMIS, OAIS, TRAC).

The Present

Current activities address digital preservation issues at three levels, namely: fundamental research, applied research & development and networking.

The current *fundamental research* moves beyond the preservation of simple documents and data structures. The focus is on interactive objects, embedded objects, ontologies and ephemeral data. An example for this development is the LIWA project addressing Web Archiving. Even more ambitious goals has the just-started project TIMBUS with research on preservation of business processes. Moreover, fundamental research is carried out on formal methods for object validation within the PLANETS and SCAPE projects. This includes the validation of objects according to format specifications and policies as well as results of preservation actions against completeness and correctness.

Applied research and development in digital preservation focuses on scalable preservation systems. The need stems from the user communities requesting tools, methods and models that perform on realistic heterogeneous large collections of complex digital objects. A second aspect of handling vast amounts of objects effectively is the automation and decision support in a number of stages, ranging from object selection, tool performance, to validation criteria. In the past a number of conceptually well designed modules for digital preservation tasks were developed that required human intervention. Current research is focused on taking these modules to the next level and providing a high degree of automation of preservation processes as well as assist decision making. Examples are the SCAPE project that is primarily addressing the scalability issue and ARCOMEM that is using the social web for automated information creation and supported appraisal. The ENSURE project will research on scalable pay-as-you-go infrastructure for preservation services for integration into workflows.

The third issue addressed by current projects is *networking*. An achievement of past projects with intensive outreaching and publication activities is the broadening of the digital preservation community. Awareness about digital preservation stretches far beyond the traditional archive, library and museum sector (ALM), now reaching the academic sector as well as the industry and enterprise domains. This development is well reflected in current project consortia with increasing participation of industry players as solution providers as well as problem owners. The increasing need for digital preservation experts can only be partially fulfilled by the staff of former research projects. An ongoing activity is the establishment of common training and education programmes addressing the interdisciplinary challenge of digital preservation. Formal training courses are required for the wider community that address the demands of different stakeholders and institutions, e.g. data producer, data processor, the academic sector as well as the industrial sector. In addition formal qualifications for professional education (e.g. university curricula) across Europe are required to ensure sustainable professional education in this area. Training as well as professional education fosters the raise of public and political awareness of the urgency of the preservation issue. Another important networking activity is the establishment of an audit and certification process. The contributions by CASPAR to an international Audit and Certification standard should now be finalized for common audit procedures and taken to the next level (e.g. APARSEN). A common approach will result in re-shaping the digital

preservation environment and establishing a stabilized landscape.

The Future

The vision about the future of digital preservation and required activities are outlined in a number of research roadmaps. One of the key observations from these is a slow shift from addressing questions that help to fix problems in maintaining digital information over time to ensuring that the problem will not appear in its full complexity in the first place, reducing the need for specific ex-post fixing. With the progress made in DP research so far, the community has developed a solid understanding of the problems and the approaches needed to fix them, turning DP activities in some areas into a challenging engineering task that requires further attention. Beyond that, however, more fundamental research is required in order to ensure that the way information and information processing systems are produced in the future pose less of a challenge in terms of preservation.

This can be seen in research challenges focussing on the development of DP-ready systems, integrating DP requirements in any system design and development process. A higher level of resiliency against technological changes on all levels will not only make preservation easier, it will also offer benefits in the operations of information systems.

A further area of focus is automation on all levels to be able to deal with the increasing amounts as well as growing levels of complexity of objects that have to be dealt with. While the focus of the former will be on scalable architectures, the focus of the latter will need to involve a more solid understanding of the fundamental concepts of digital information including entire systems and distributed processes.

We also observe a shift in the community recognizing the need for preservation solutions and thus also stakeholders in DP related research and development. While originally being strongly based in the cultural heritage and scientific data domain, stakeholders from a range of other disciplines involved in e-* activities (e-health, e-government, e-commerce) realize their dependency on electronic information and processes beyond legal retention requirements for their very operations. This will have an impact on the type of solutions expected, as well as the approaches taken to meet these, broadening both the interdisciplinarity as well as the methodological approaches to be taken.

With digital preservation having evolved into a dedicated and highly specialized discipline in its own right, a further challenge now will be to reach out again to other disciplines to bring in know-how from highly specialized domains. Within the ICT domain, this will require attracting input from groups e.g in the area of HW and embedded systems design, algorithm and compilers, theory of computing, security, semantic technologies, to software engineering and enterprise architectures and many others. To address the technological challenges in digital preservation specifically within the broadening application domains where solutions are needed will require teams integrating experts from a range of ICT disciplines, organizational and legal experts and domain experts to cover

the entire lifecycle and operational context of an information system.

This growing need for expertise in an increasing market will also call for a broader level of educational offering. This will require both fundamental education to further the field as well as solid training to actually manage the preservation of the information and processes.

In a nutshell, digital preservation research and development has advanced impressively. It has evolved into a large community of experts, developed a solid understanding of the problems to master, and developed solutions that help to address the challenges faced by current stakeholders. Significant efforts will be required to proactively address emerging challenges at new levels of scale and complexities on a foundational level as virtually all areas of society are starting to face preservation challenges with processes depending on ubiquitous information technology.

Contents

1	Introduction	8
1.1	The Problem of digital preservation	8
1.2	Projects overview	8
1.2.1	Key figures	8
1.2.2	Digital Preservation Projects in the ICT program	9
1.2.3	Relationships between projects	11
1.3	Structure Of This Report	12
2	Research projects	13
2.1	Content Types	13
2.2	Targeted Audience	14
2.3	Research partners	15
2.4	Core Objectives	16
2.4.1	Focused topics, broader application and new approaches	17
2.4.2	Scalability	18
2.4.3	Intelligent tools and approaches	18
2.4.4	Conceptual models & system design	18
2.4.5	Authenticity, Trust, Audit	19
2.4.6	Metadata	20
2.4.7	Semantic technologies	20
2.5	Training & Education	21
2.6	Related disciplines	21
3	Research Agendas	24
3.1	DPE research agenda	24
3.2	Roadmap of PARSE.Insight	25
3.3	Dagstuhl Seminar on automation in digital preservation	26
3.4	DP Research Challenges Wiki	27
3.4.1	Focus of infrastructure-based research	28
3.4.2	Focus of content-based research	29
4	Conclusion	30
5	Appendix	31
5.1	APARSEN	32
5.2	ARCOMEM	33
5.3	BLOGFOREVER	34
5.4	CASPAR	35
5.5	DELOS	36
5.6	DPE	37
5.7	ENSURE	38
5.8	ERPANET	39
5.9	KEEP	40
5.10	LiWA	41

5.11	PARSE.Insight	42
5.12	PLANETS	43
5.13	PrestoPRIME	44
5.14	PROTAGE	45
5.15	SCAPE	46
5.16	SHAMAN	47
5.17	TIMBUS	48
5.18	WF4EVER	49

1 Introduction

In the last 30 years information technology has changed the way of how we think about, create, store, represent and share information more than ever before. This rapid and drastic change leads to many improvements, inventions and discoveries of things unimaginable before, but leaves us with a great problem in terms of preserving the digital data, the newly gained knowledge and our cultural heritage.

This section will provide a short overview of the domain and the projects analyzed for this report.

1.1 The Problem of digital preservation

30 years ago almost all content was created on paper. Media and people were able to store it for long periods of time under the right conditions. Today's media does not work as effective as paper in the long term. Rapid changes in technology lead to the invention of faster hardware, more efficient and robust software and a great deal of new format specifications. Thus the content stored on old media, rendered with old software and formatted with old specifications often does not stand a chance to live through a single decade.

Fortunately, the European Commission has recognized this problem at an early stage, even though it often tends to be neglected by business and industry. To prevent the loss of data, a number of research projects concentrating on preservation of digital content are partially funded throughout the EU framework programmes for research and technological development.

1.2 Projects overview

The following section provides an overview of projects related to digital preservation in the 6th and the 7th framework programmes of the European Union and discusses project relationships and possible future work.

1.2.1 Key figures

Altogether over 90 million Euro are being invested in digital preservation related topics (not exclusively) in both programmes, divided amongst more than 15 projects focused on digital preservation with partners from more than 20 countries. The 6th framework programme started in 2002 and lasted till 2006, with FP7 running from 2007 till 2013.

Figure 1 gives an overview of the funding on a project basis. It is important to note, that the funding of digital preservation projects in the seventh framework programme has more than tripled, which indicates the awareness of the problem. With the increase in financial support not only the number of research projects but also the group of stake-holders and institutions active in digital preservation has risen, leading to a solid body of DP expertise across a range of domains and institutional backgrounds.

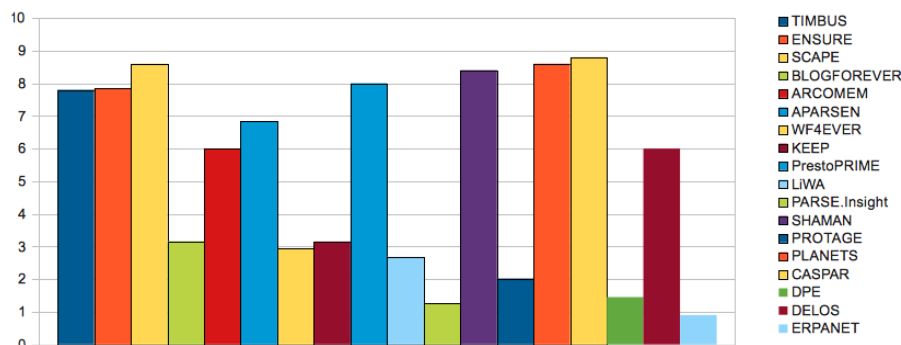


Figure 1: EU funding for all digital preservation projects in (mio) EUR

1.2.2 Digital Preservation Projects in the ICT program

The following set of projects forms the basis of the evaluation provided in Section 2. The short descriptions are mostly adapted from the various projects' websites. A detailed listing giving key characteristics is provided in the Appendix.

Current projects (in reverse order of starting date)

TIMBUS (FP7, IP) will endeavor to enlarge the understanding of DP to include the set of activities, processes and tools that ensure continued access to services and software necessary to produce the context within which information can be accessed, properly rendered, validated and transformed into knowledge.

Wf4Ever (FP7, STREP) aims at providing the methods and tools required to ensure the long-term preservation of scientific workflows.

ENSURE (FP7, IP) will ensure the long term usability for the spiraling amounts of data produced or controlled by organizations with commercial interests. It will significantly extend the state of the art in digital preservation which to-date has focused on relatively homogeneous cultural heritage data through analyzing use cases from diverse fields such as health care, clinical studies, and financial services.

SCAPE (FP7, IP) will enhance the state of the art of digital preservation in three ways: by developing an infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system.

BlogForever (STREP) will create digital preservation, management and dissemination facilities for weblogs.

ARCOMEM (FP7, IP) The vision of the ARCOMEM project is to leverage the Wisdom of the Crowds for content appraisal, selection and preservation, so that archives reflect collective memory and social content.

APARSEN (FP7, NoE) 'Alliance Permanent Access to the Records of Science in Europe' - is a Network of Excellence gathering digital preservation practitioners and researchers.

KEEP (FP7, IP)) is developing emulation services to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames, etc. The technology is being tested on early computer games.

PrestoPRIME (FP7, IP) is addressing long-term preservation of and access to digital audio-visual content by integrating media archives with European on-line digital libraries. Research will result in a range of tools and services, delivered through the networked Competence Centre PrestoCentre.

LiWA (FP7, STREP) develops and demonstrates web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long term interpretability of web content.

SHAMAN (FP7, IP) is developing a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives.

Past projects

PARSE.Insight (FP7, CA) aims to highlight the longevity and vulnerability of digital research data and concentrates on the parts of e-Science infrastructure needed to support persistence and understandability of the digital assets of EU research.

PROTAGE (FP7, STREP) stands for Preservation Organizations Using Tools in Agent Environments - it addressed the challenges related to the preservation of digital resources of increasing volume and heterogeneity by developing tools allowing for more efficiency and self-reliance of preservation processes.

PLANETS (FP6, IP) stands for Preservation and Long-term Access to our Cultural and Scientific Heritage. Its primary goal was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets. The project delivered an integrated production environment for the management of digital information preservation, with a special focus on the needs of libraries and archives.

CASPAR (FP6, IP) stands for Cultural, Artistic and Scientific Knowledge Preservation, for Access and Retrieval. The CASPAR team created a

framework of tools and infrastructure components to support the end-to-end preservation of all types of digitally encoded information and thus help producers, curators and users of digital resources share the burden of preservation.

DPE - DigitalPreservationEurope (FP6, CA) was a coordination action project and was launched in order to improve cooperation and consistency in current activities to secure effective preservation of digital materials. The project has led work to raise the profile of digital preservation; to promote auditable and certified standards for digital preservation processes; and to facilitate skills development through training.

DELOS (FP6, NoE) was a Network of Excellence on Digital Libraries. It carried out research in the fields of library architectures, information access and personalisation, audio-visual and non traditional objects, user interfaces, knowledge extraction, semantic interoperability, preservation and evaluation.

ERPANET (FP5) aimed at establishing an expandable and self-sustaining European Initiative, which serves as a virtual clearinghouse and knowledge-base in the area of preservation of cultural heritage and scientific digital objects.

1.2.3 Relationships between projects

A number of project relationships exists in the digital preservation area. The continuous work of these projects evidences the sustainable research that happens in this area. A Gantt Chart of the projects can be seen in Figure 2 where the colour blue stands for a 'Specific Targeted Research Project' project type, red for 'Network of Excellence', yellow for 'Coordinated Action' and green for 'Integrated Project'.

The DELOS Digital Preservation Cluster (DPC) built on the earlier successful work of ERPANET. It continued and extended the work of ERPANET, key players of the project consortium remained the same. DigitalPreservationEurope(DPE) refers to work of ERPANET and the DELOS DPC. The project continued and extended the work on brief guidelines and overviews of key topics. The evaluation framework of the DELOS project built the basis for the PLANETS Planning approach. This work will be further explored within the SCAPE project.

The recently started APARSEN NoE refers to work of the PARSE-Insight project such as the roadmap and surveys and also to work of CASPAR (authenticity tools).

Another successful take-up of previous project results and continuous work comes from the audiovisual sector with PRESTO¹. PrestoSpace² extended the

¹<http://presto.joanneum.ac.at>

²<http://prestospace.org>

approaches of PRESTO outside the broadcasting setting. Both projects developed preservation factory approaches and methods for digitizing audiovisual collections. The latest project PrestoPRIME builds on the work of PRESTO and PrestoSpace with a focus on +digital preservation of audiovisual content. The preservation of 'born digital' audiovisual content is addressed by PrestoPRIME. In March 2011 the networked competence centre PrestoCentre³ was launched. PrestoCentre is a membership driven organisation that brings together a community of stakeholders in audiovisual digitisation and digital preservation. Similarly, the Open Planets Foundation⁴ (OPF) was created as a membership-driven successor of the PLANETS project to provide survivability to the solutions developed in the project.

The continuation of the projects as well as the creation of larger membership organisations to carry forward the results obtained indicate the good networking within the area in digital preservation. The consistency of research activities in this area leads to the establishment of competence centers with a wealth of experience and excellent expert knowledge within Europe. An analysis of project partners can be found in Section 2.3.

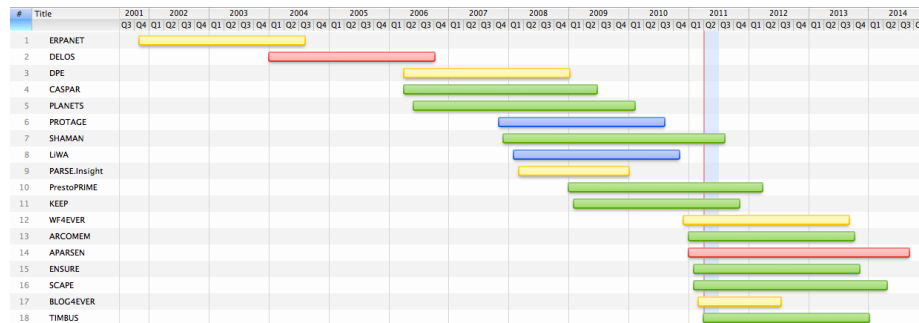


Figure 2: Timeline of the ICT Projects

1.3 Structure Of This Report

The remainder of this report is structured as follows; Section 2 summarises different aspects of the projects and thus provides new perspectives on the research on digital preservation in Europe. Section 3 presents three research agendas proposed and intrudes the research challenge wiki. In Section 4 a conclusion on the work of the reach projects is drawn. The Appendix provides key factors on all projects discussed in this report.

³<http://www.prestocentre.eu>

⁴<http://www.openplanetsfoundation.org>

2 Research projects

In this section the projects are categorized based on different criteria, such as content type discussed, core objectives and targeted audiences, in order to give a better overview of their commonalities and to provide a new perspective on their relations. The projects discussed here have their main focus in digital preservation or had an important impact in the field. Other projects, supporting digital preservation or related to it, but not having it as a force focus area (such as digitization projects) are not discussed here.

2.1 Content Types

In this section the projects are analysed according to the types of objects addressed. The list does not pretend to be complete and only represents the content types that are specifically addressed by the projects. It does not mean that other content types are not covered or supported by the projects outcomes. The categories that emerged while analysing the projects were aggregated into eight main categories:

Office Documents (including all kinds of text documents and images), audio/visual content, scientific data, web content, social web, interactive content, applications and processes. No content type was assigned to the CA DPE as no focused research on specific content types was done.

The projects in Table 3 are sorted by their starting year. It is clear that a shift in the analyzed content types occurs. The projects from FP5 and FP6 concentrated on office documents, including images in institutional settings. The projects took a broad approach in respect to the content, aiming at providing basic concepts and tools that can be used in many settings. Some work was already done on the preservation of audio-visual and scientific data by projects such as PLANETS, CASPAR and SHAMAN. SHAMAN was the first project that explicitly addressed the requirements of product life cycle management (PLM) and workflows that are relevant for preservation in the design- and engineering domain.

However, the real shift to more complex data structures and formats took place in the seventh framework programme. Projects like LiWA, SCAPE, BLOGFOREVER and ARCOMEM deal with scientific, social and web content. They are tackling different aspects of the problem of preserving generic documents. ENSURE deals with data from healthcare and financial sector.

A development can be identified from Table 3. The content types can be seen to become more complex from left to right. The early projects concentrated on single documents such as office documents or audiovisual content. The table shows a shift to the right for the new projects. SCAPE and ENSURE are the exception here, they deal with large scale preservation of office documents (SCAPE) to clinical trials (ENSURE).

The current focus of research is on interactive objects, embedded objects, ontologies and ephemeral data. This shift becomes even more evident when analyzing the digital material that will be addressed by the just-starting F7 IPs

TIMBUS and WF4EVER. The TIMBUS project researches on the support for preservation of business processes and applications. WF4EVER investigates a technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows. Both projects represent a shift in paradigms allowing to see the data, structure and behavior detached from physical files. It allows a system wide view on the actual data, leaving behind the file centric view of systems.

Project/ ContentType	Starting Year	Office Documents	Audio/Visual	Scientific data	Web Content	Social Web Content	Interactive Content	Applications	Processes
TIMBUS	2011						x	x	
ENSURE	2011		x						
SCAPE	2011	x	x	x					
BLOGFOREVER	2011				x	x	x		
ARCOMEM	2011				x	x			
APARSEN	2011		x						
WF4EVER	2010		x					x	
KEEP	2009						x	x	
PrestoPRIME	2009		x						
LiWA	2008				x				
PARSE.Insight	2008	x		x					
SHAMAN	2007	x	x	x				x	
PROTAGE	2007	x							
PLANETS	2006	x	x						
CASPAR	2006	x	x	x	x			x	
DPE	2006								
DELOS	2004	x							
ERPANET	2001	x		x	x				

Figure 3: Focused content types of the projects

2.2 Targeted Audience

Table 4 lists the targeted audiences for each project, providing another perspective on the research concentration. The table divides the targets into 5 different categories: Memory Institutions, Scientific Institutions, Government Organizations, Enterprises, and Private. Memory Institutions not only include libraries and archives, but also online web archives, digital libraries and digital repositories. The Private category sums up the end user consumer that cannot be categorized in any other field.

Table 4 shows a concentration on memory institutions and scientific institutions. This is not surprising as the museums, archives and libraries were the first group facing the problem of digital preservation. They have the obligation to preserve their holdings (analogue and digital) for the long term. Therefore, many key players in the area of digital preservation are memory institutions.

SHAMAN and PROTAGE where the first projects that addresses enterprises with their research.

A series of projects have not identified a special target audience group as they researched on generic issues and basic technologies that are generally applicable.

The just-started research projects (TIMBUS, ENSURE and SCAPE) explicitly shift their focus to the need of the business sector. TIMBUS will focus on the preservation of business processes, while SCAPE will develop scalable services and a platform for orchestration of semi-automated workflows for large-scale, heterogeneous collections. ENSURE will research on scalable pay-as-you-go infrastructure for preservation considering economic implications.

Project/ Target Institutions	Starting Year	Memory Institutions	Scientific Institutions	Government Organizations	Enterprise	Private
TIMBUS	2011		x		x	
ENSURE	2011				x	x
SCAPE	2011	x	x		x	
BLOGFOREVER	2011	x	x		x	x
ARCOMEM	2011	x				
APARSEN	2011	x	x			x
WF4EVER	2010		x			
KEEP	2009	x				x
PrestoPRIME	2009	x	x			
LIWA	2008	x				
PARSE.Insight	2008		x			
SHAMAN	2007	x	x	x	x	
PROTAGE	2007				x	
PLANETS	2006	x	x	x		
CASPAR	2006	x	x	x		
DPE	2006	x	x	x		
DELOS	2004	x				
ERPANET	2001	x	x	x		

Figure 4: The projects' targeted audiences.

2.3 Research partners

Table 5 shows the list of partners that were involved in two or more ICT projects about digital preservation. The first column states whether it is a heritage institution (ALM), a scientific intuition (SCI) or an industry partner (IND).

It is not surprising that a number of large national libraries is on the list. In the previous projects only two national archives were involved in more than two projects (Netherlands and Switzerland). While the libraries were active in the past, the list shows only two new projects with libraries involved (APARSEN and SCAPE). This trend is also shown in Section 2.1 and 2.2 with target audience and content type. We observe a shift of target content types from rather simple documents (that were held in large quantities by libraries) to more complex data and processes. The target audience also shifts towards broader sectors

of industry and business (see Section 2.2). Table 5 shows a large number of scientific institutions that were involved in digital preservation research projects. A number of universities from Germany and the United Kingdom research on this topic. The large number of research institutions that were involved in two or more research projects indicate the continued work within their intuitions on digital preservation.

All industry partners listed in Table 5 offer solutions and products for digital preservation as their business. While they were so far only partners in the consortium, with TIMBUS and ENSURE there are two projects lead by an industry partner, where they also represent problem owners rather than solely solution providers.

Partner Type	Partner Institution	Country	ERRANET	DELOS	DPE	CASPAR	PLANETS	PROTAGE	SHAMAN	LWA	PARSE: Insight	PrestoPRIME	KEEP	APARSEN	ARCOMEM	BLOGFOREVER	SCAPE	ENSURE	WEFAEVER	TIMBUS
ALM	The British Library	United Kingdom					x							x			x			
ALM	Deutsche Nationalbibliothek	Germany							x		x		x	x						
ALM	Koninklijke Bibliotheek	Netherlands									x		x	x			x			
ALM	Narodni Knihovna Ceske Republiky	Czech republic			x					x										
ALM	Nationaal Archief	Netherlands	x	x	x		x													
ALM	Österreichische Nationalbibliothek	Austria		x			x							x				x		
ALM	Schweizerisches Bundesarchiv	Switzerland	x				x													
ALM	Statsbiblioteket	Denmark			x		x												x	
SCI	Austrian Institute of Technology	Austria					x												x	
SCI	Aristotle University of Thessaloniki	Greece		x													x			
SCI	Digital Preservation Coalition	United Kingdom												x						x
SCI	European Organization for Nuclear Research	Switzerland									x		x	x		x				
SCI	European Space Agency	France				x					x		x							
SCI	Fernuniversität in Hagen	Germany			x				x		x									
SCI	Fondazione Rinascimento Digitale	Italy		x	x															
SCI	Foundation For Research And Technology Hellas	Greece				x								x						
SCI	INESC ID	Portugal							x											x
SCI	Institut National de l'Audiovisuel	France			x						x									
SCI	Internet Memory Foundation	Netherlands							x						x		x			
SCI	L3S Research Center	Germany							x						x					
SCI	Max Planck Gesellschaft	Germany		x						x	x									
SCI	Science and Technology Facilities Council	United Kingdom				x					x			x			x	x		
SCI	Stichting Nederlands Instituut voor Beeld en Geluid	Netherlands								x		x								
SCI	Technische Universität Berlin	Germany															x	x		
SCI	Technische Universität Wien	Austria		x	x		x											x		
SCI	The University of Glasgow	United Kingdom	x	x	x	x	x		x								x			
SCI	University of Liverpool	United Kingdom							x			x		x						
SCI	The University of Manchester	United Kingdom														x	x		x	
SCI	Università di Urbino	Italy	x	x		x														
SCI	University Of Southampton	United Kingdom		x								x			x					
SCI	Secure Business Austria	Austria												x						x
SCI	Staats- und Universitätsbibliothek Göttingen	Germany		x	x				x		x									
IND	Ex Libris LTD.	Israel										x						x		
IND	IBM – Science And Technology LTD	Israel				x								x					x	
IND	Micrisoft Research Limited	United Kingdom					x								x				x	
IND	Philips	Netherlands							x											x
IND	Tessella Support Services PLC	United Kingdom					x						x	x						x

Figure 5: Institutions that were in two or more project consortium

2.4 Core Objectives

In this section core objectives of the research projects are identified. The consolidation was done based on the project descriptions, the call texts and research

roadmaps.

Project/ Objectives	ERPANET	DELOS	DPE	CASPAR	PLANETS	PROTAGE	SHAMAN	LIWA	PARSE Insight	PrestoPRIME	KEEP	APARSEN	ARCOMEM	BLOGFOREVER	SCAPE	ENSURE	WF4EVER	TIMBUS
Appraisal and Selection	X							X					X					
Characterization				X	X		X			X					X			
Format Identification				X						X								
Metadata				X			X	X		X	X		X					
Preservation Action				X	X					X	X	X			X			
Preservation Planning		X			X					X					X			
Authenticity & Trust			X	X		X		X	X	X						X		X
Access		X		X			X		X	X	X							
System design				X		X	X			X							X	X
Workflows				X	X										X	X	X	X
Tool Development				X	X	X	X			X			X	X			X	
Interoperability				X	X									X				
Scalability							X	X		X					X	X		
Legal			X							X	X	X		X				X
Research Roadmap			X						X						X			
Training	X	X	X	X	X							X						
Coordination			X									X						

Figure 6: Core objectives of the projects

2.4.1 Focused topics, broader application and new approaches

The general development of DP research in Europe can be portrayed well by the short overview of core topics and aims of the projects. The early research projects in the field of digital preservation started with the definition, design and discussion of basic concepts, systems and methods. The projects were mainly driven by the digital library and archive community. The dominant topics included, amongst others, selection and appraisal, metadata definitions, unique identifier, characterization tools. Projects of the first phase were ERPANET and DELOS.

In a next stage, available methods, tools and modules were integrated into framework architectures. The integration of digital preservation modules into framework architectures allows the composition of workflows and integration into other system. It fosters the broader application of DP tools. Examples are the PLANETS Interoperability Framework, the CASPAR Integrated Framework and the integrated preservation framework using grid-technologies of SHAMAN. An overview about the frameworks is provided in [12].

In a next round of projects more specialized application scenarios and new approaches were addressed. The tools and methods developed so far were focused on boread application scenarios. The new approaches include the agent environment of the PROTAGE project or the use of Social Web in the ARCOMEM project. Focused and specialized topics are addressed both within the LiWA project with web archiving and the KEEP project with emulation that was only discussed marginally by other research projects so far.

2.4.2 Scalability

Existing tools developed for digital preservation were mostly developed to demonstrate specific functionalities. They were not designed to operate at large scale. In practice we are facing sheer volumes of content, e.g. web archives or repositories of larger institutions (such as archives or libraries). There is a need of the community for scalable tools and methods that are able to process a large number of objects. SCAPE, ENSURE and LIWA are addressing the scalability of preservation solutions. TIMBUS and ENSURE are researching on cloud storage for scalability, as well as the use of virtualization technologies for preservation.

2.4.3 Intelligent tools and approaches

The first tools developed required human interaction and profound knowledge of digital preservation. The current trend is towards intelligent tools and approaches that assist the users and support the decision making process. By creating and using knowledge bases and innovative approaches, tools can move to the next level of supporting complex settings with a high volume and heterogeneous content.

An example is the Plato preservation planning approach. The fundamental concept has been developed within the DELOS project, it was refined based on practical experience and feedback in several phases in the PLANETS project. It results in a systematic approach for well-documented, well-argued and transparent decisions supported by a software implementation with integrated knowledge base, semi-automated service discovery and automated measurements and comparison of original and migrated objects. The work on Plato will be continued within SCAPE focusing on large scale content and providing a higher degree of automation for the decision making process.

Another example is GRATE (Global Remote Access To Emulation) developed within PLANETS that allows the wrapping of different rendering environments over a remote network. It eliminates the need of local installations of different emulators.

The novel PROTAGE approach to digital preservation is a flexible and distributed software agent system. In this system agents work autonomously. They need to make decisions, perform preservation tasks and collaborate with other resources. The system supports and aids users in preserving, retrieving and sharing digital objects.

2.4.4 Conceptual models & system design

The OAIS reference model and its terminology has been established as a common basis for concepts and models in the area of digital preservation. Work on conceptual models has been done by a series of research projects.

The CASPAR [3] conceptual model is strongly influenced by the OAIS model. It provides a generic infrastructure concept to support digital preservation, where key preservation components were identified. The model covers

also an information model, discussing in detail the concept of Representation Information and Preservation Description Information.

First thoughts about how to address digital preservation from a system design perspective have been done by the SHAMAN project [2]. It identified the main characteristics and requirements of systems and motivated the use of Enterprise Architecture Frameworks to address digital preservation. SHAMAN designed a reference architecture based on the OAIS Reference Model. The SHAMAN Reference Architecture [14] provides several viewpoints that reflect the concerns of the stakeholders. The model provides a process that should help to derive concrete digital preservation architectures for a targeted environment.

A conceptual model [7] capturing requirements was developed within the PLANETS projects. It links risks with the actions that mitigate them and expresses them in stakeholder specific requirements.

A different approach of system design was used by PROTAGE using a Multi Agent System Architecture. The architecture consists of software agent tools and web services for long-term digital preservation and access. Agents coordinate the preservation process, they automatically locate, select and employ web services to obtain information, to make decisions and to perform the preservation tasks. Web services are a set of services that provide particular types of services for the agents and the user, such as access, serve as a knowledge base, virus checks, migration and metadata extraction.

TIMBUS will investigate on reference architectures for intelligent Enterprise Risk Management systems with preservation functionalities. In the project the system design aspect will be addressed in more detail, e.g. providing guidelines for developing software services and systems which are digitally preservable.

Workflows for preservation are addressed by SCAPE and ENSURE. The project researches on the design and orchestration of preservation workflows for digital objects.

2.4.5 Authenticity, Trust, Audit

A common requirement of all digital preservation systems is the authenticity of digital objects. This includes to assure the integrity of digital objects, guaranteeing that their informational content was not modified. The current effort in the field of authenticity of objects is the establishment of common understanding and a theoretic basis for the concept of authenticity and integrity. For the technical aspect of authenticity, the research projects are developing approaches for different types of objects.

Very little research has been done on the preservation of the semantic level of digital objects as part of authenticity. The LiWA project addressed the changes in language over time as part of semantic preservation. In order to interpret the content on the long run, they have been developing methods for automatically dealing with terminology evolution.

Trust is an important aspect for long term archives. DPE was the first project that addressed this issue through the development of a self assessment method (DRAMBORA). It encourages organizations to establish a comprehen-

sive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organization.

Work on trust and authenticity with respect to digital repositories has been done within the CASPAR project. The project contributed to the update of the treatment of authenticity and DRM in OAIS. It was successful in contributing to the development of the ISO audit and certification process for digital preservation. Within APARSEN common methods for independent 3rd party certification of repositories will be developed.

In the context of certification legal issues of digital preservation are of interest. They were left out by many projects. DPE did some initial work in this area and KEEP published legal studies about emulation. CASPAR did some work on DRM and developed a tool to deal with a changing legislative base and accommodate a multiplicity of legal systems. TIMBUS will address the the legal issues of preserving business processes.

2.4.6 Metadata

A lot of effort was invested in metadata for digital preservation [5, 8, 1, 10]. A number of standards for descriptive metadata have been set up that are widely accepted and in use, e.g. MAB, MARC, Dublin Core. For digital preservation PREMIS has been established as core preservation metadata standard.

While communities have a strong desire for long-lasting, stable metadata standards, initiatives continue to evolve and extend metadata standards and schemata. The motivation behind new standards, recommendation and implementation is the better support and coverage of different settings, aims, objects and approaches.

Work on metadata has been done by almost all research projects, CASPAR researched on Descriptive Information for objects and on Representation Information. PLANETS was investigating on advanced characteristics and created a data dictionary for key digital preservation metadata concepts [6]. The underlying conceptual model supports dynamic preservation processes, rather than the static recording of characteristics and events.

Application-specific research on metadata is conducted by KEEP, LiWA, ARCOMEM and PrestoPRIME for supporting emulators, web archiving, social web and audiovisual content.

2.4.7 Semantic technologies

The CASPAR project used semantic web technologies to model knowledge for digital preservation. The knowledge management service contains information about Representation Information and the Knowledge Bases of Designated Communities.

The SHAMAN context model provides an infrastructure-independent representation of the attributes and relations between digital objects by using ontologies.

2.5 Training & Education

Almost all research projects offer training courses about their work in digital preservation. Starting with training seminars of ERPANET and summer schools offered by the DELOS DPC, each project organizes public events giving introduction to DP and their achievements. First consolidation work of training was done by DPE, PLANETS and CASPAR with the wePreserve platform. They offered joint training events about their work and basic concepts of digital preservation. This principle could not attract imitators and the training offers are again very fragmented. At present formal qualifications for professional training are missing. There is a need from different sectors (academic, industry, culture heritage, private sectors) for continuous professional training. The training should address technical as well as organizational aspects of digital preservation.

In addition to the formal training, a professional education program for digital preservation is required. The education program should overcome the current shortage of librarian and digital curator education addressing the interdisciplinary aspects of digital preservation.

The APARSEN project will address the consolidation of training courses and material which will allow the creation of coherent training courses and formal qualifications.

2.6 Related disciplines

Digital preservation has established itself as independent scientific discipline in the recent years. The inter-disciplinarity of DP has many relations and overlaps to other disciplines and topics. In this section, pointers to a selection of other disciplines are outlined. These are necessarily incomplete and should be seen as initiators of a broader discussion that is required to clarify boundaries and relationships. In this list the disciplines where digital preservation has been developed from, such as areas digital libraries and archives, are left out.

Information Management, Data Management, Knowledge Management

Digital preservation is sometimes coined 'Information Management with a long-term perspective', and as such closely connected with the named areas. Efficient data management is required for large-scale preservation, and knowledge management approaches may address in particular the preservation of semantics of information.

Information retrieval & Data mining

Information retrieval and data mining provide essential techniques to handle large volumes of data. They can support classification of documents and terms within documents in an efficient way. Appraisal approaches, for example, can benefit from support of this area.

Semantic Technologies

Too little attention is paid to the semantic layer of digital objects in current approaches of digital preservation. Semantic Technologies such as ontologies can help to preserve the meaning of data in the long run. It will support access on the data and understanding of the data by the user. Other approaches such as reasoning can help to process and prepare the existing data for archiving.

Enterprise Architecture

For many organizations, digital preservation is becoming a fundamental requirement in order to enable delivery of information and processes in usable forms across and outside the enterprise. This requires information, services and technology to be aligned with organizational structures and business goals, which is a core aspect in Enterprise Architecture.

Service orientation and cloud computing

Paradigms such as SOAs and cloud computing provide new threats and opportunities for digital preservation. They are implementing new levels of abstraction and provide modular system architectures that can ease, but also complicate the preservation of systems. The increasing distribution of systems and use of web services increases the complexity of systems and can cause difficulties in understanding and documenting processes in these systems as a whole.

Databases

Databases are key components of all information system containing structured organized data. Current systems provide an extensive range of functionality and are reaching a high degree of complexity. Digital preservation capabilities are required in the system to ensure the long term access of the data and the embedded functionalities.

Storage technology

Current storage technology is optimized for aspects such as access speed and throughput, but not for long term storage of data. For DP, it would be attractive if it was possible to dynamically decide for a suitable storage technology whether to prioritize access speed over durability of storage or vice versa. This further requires appropriate cost models for storage.

Risk management

Risk management is an elementary aspect of digital preservation. Risk management methods and concepts can help to ensure the usability and accessibility of information over time. Digital preservation systems should provide systematic risk management to mitigate the risk of data loss.

Governance, best practice, compliance and trust

Common governance and compliance guidelines are important aspects for digital preservation archives. Existing frameworks in IT-management can provide a basis for digital preservation initiatives. The establishment of further certification initiatives supports repositories to establish trusted status amongst their stakeholders.

Legislation

Currently, common legal understanding and legislation for digital preservation in Europe is missing. A harmonization of legislation is needed to ensure legal compliance of preservation approaches and systems. In order to provide legal certainty for vendors and users of DP systems a common legal framework is desirable.

Digital Rights Management

Digital preservation is touching a number of legal issues for instance intellectual property rights. Efficient Digital Rights Management (DRM) systems are required that are aware of rights & permissions regarding the digital content. Moreover, these systems need to track changes in rights over time.

Security

Digital preservation needs to guarantee information confidentiality, availability, integrity, and authentication over the long term. Time is a critical component for security mechanisms. Current techniques and methods cannot be considered as secure in a distant future due to novel attacks and increasing computation power.

For digital preservation security mechanisms such as certificates are required that take the time frame into account and support changes of technology (e.g. encryption methods). The mechanisms need nonetheless always be traceable and ensure the integrity and authentication over time.

Privacy

A large number of digital preservation scenarios are dealing with personal information (such as e-health, e-government, etc.). Ensuring the privacy needs to be a main objective for such systems. Current approaches in e-health are implementing different level of access and anonymization of personal data. These approaches need to be extended for the long run.

3 Research Agendas

The problem of digital preservation started to emerge about 20 years ago with ad-hoc initiatives addressing it. Several research agendas were subsequently created to address these issues more systematically. This section builds on top of the DPE research agenda [9], which summarized the earlier agendas, as well as the roadmap of the PARSE.Insight project [11] and the outcome of the Dagstuhl Seminar [4]. The Dagstuhl seminar attempted to create an outline of specifically the IT aspects requiring consolidated research. The PARSE.Insight project provided a roadmap for a scientific data infrastructure identifying missing components, both technical and non-technical. Within the Pares.Insight project a large scale survey was conducted within three stake-holder domains: research, publishing and data management. The results of the survey provided input to the research roadmap. In this report we list only the research topics, skipping solely organizational, engineering and development issues.

3.1 DPE research agenda

DPE carried out a thorough analysis of all existing research agendas in order to sum up what had to be done and identify missing aspects of the problem which ultimately led to the foundation of Europe's research and development in terms of digital preservation. Altogether 10 fields of research were proposed:

Restoration - even though there are forensic methods to physically restore data from damaged media the rendering of these objects is a problem considering the fact that their type is unknown. Thus, the rendering of these objects and revealing their actual content forms a significant challenge for digital preservation.

Conservation - in order to cope with obsolescence of technology, methods and strategies such as migration and emulation have emerged, providing a way to continuously preserve older data. Nevertheless, these introduce further challenges and reveal new topics for research and development.

Management - Research needs to focus on the planning, enacting, executing, managing and monitoring of organizational processes for digital preservation.

Risk - Basically, digital preservation can be seen as a risk problem. In the end it all comes down to making a choice and deciding which alternative is the most appropriate considering many (uncertain) factors, such as organization policies, data collections, costs, etc. Thus decision making tools and instruments that automate the solution of these problems are required.

Significant Properties of Digital Objects - these are required in order to understand an object and keep it usable and authentic in the long-term. Thus research in capturing but also in preserving these properties and their

relations was needed. Moreover significant properties allow measurements for preservation actions.

Interoperability - a great deal of formats and format types exist already and new emerge every day. All present solutions specialize only in a subset of these. There are repositories that could handle any kind of digitally encoded data, however organizations and institutions are often forced to use a number of solutions. Therefore, interoperability and trust between these different service providers is essential for digital preservation.

Automation - the process of preserving digital data consists of a number of steps, which are often executed manually and the results are often aggregated and given as an input of a subsequent step. All that strongly suggests the need of automation of the process.

Context - even though digital objects tend to carry information that describes a single aspect of an issue the context and environment in which they were created plays an important role for long-term preservation. Knowing the context, the policies of organization, relations to other objects, etc. is essential for the understanding of that object in time.

Storage - Despite the methods used for preserving digital data, despite the efforts made in optimizing the size of the data, the problem with insufficient storage will always persist. However, this field plays an important role for research as on Grid and similar technologies.

Experimentation - As in every science, field experimentation is the only way to help understand users' interactions and needs with digital repositories and thus designing relevant testbeds and experiments will play an important role for digital preservation. Experimentation will leads to solid, grounded engineering and the development of evidence-based methods.

3.2 Roadmap of PARSE.Insight

The PARSE.Insight roadmap provides an overview of components and aspects that need to be available for long term preservation infrastructures for scientific data.

Financial infrastructure A lack of concepts and components for financial aspects in digital preservation were identified. Business models for digital preservation are different to other scenarios because of the long term aspect. Business models and funding schemes for services and components need to be established for a stable, robust and scalable infrastructure (e.g. storage facilities).

Virtualisation of policies, resource and processes Virtualisation is a commonly used technique to insulate services from the underlying implementation. It raises a number of requirements including

- standards for interpretability between services
- abstraction of services (such as storage)
- replication of storage resources
- logical namespaces for resources, data and users

Shared knowledge about representation information An enhanced representation knowledge management can provide a semi-automated way to verify Representation Information or provide adequate information. A shared knowledge base for Representation Information could provide a basis for such a service. Automated capturing of the creation and processing context can help to automate the preservation process.

Shared knowledge about hardware and software The aim is a set of services which make it easier to exchange information about obsolescence of hardware and software and techniques for overcoming these.

Authenticity of digital objects In order to provide evidences for the authenticity of digital objects, common formalism, standard and policies are required. They should allow a user in the future to judge the degree of authenticity which may be attributed to a digital object.

Digital Rights The aim is the ability to deal with digital rights correctly in a changing and evolving environment There are several legal systems that are subject to continuous change. Digital rights management raises a number of research issues for preservation planning and preservation actions.

Certification of repositories The aim is an audit and certification process for digital preservation systems with appropriate tools and best practice guides.

3.3 Dagstuhl Seminar on automation in digital preservation

In July 2010, there was a seminar in Schloss Dagstuhl, at the Leibniz Center for Informatics aiming to identify emerging issues in digital preservation and define the course of future research and development [4]. In the following a summary of the research issues and questions identified at the seminar is presented. Altogether they were grouped into 7 topics, which described different issues regarding digital preservation. These seem to be more generic than the ones identified throughout the DPE project, however, they cover all fields of research proposed by DPE and offer a wider range of new fields with regard to new technologies that have emerged in the meantime.

Preservation Ready Systems - the basic idea is not to build information archiving systems, which try to solve the problem but to integrate digital preservation solutions into existing systems and turn them into preservation ready systems, ultimately leading to problem prevention.

Beyond Metadata - it was identified that tools and modeling frameworks are needed. These will provide a way for capturing information about the intended and actual use of digital objects, a way to automatically establish a documentation about the digital objects and the processes that they are involved in.

Storage Technologies and Protocols - this category summarizes not only the physical storage of digital objects but actually includes a lot of innovative ideas and many research challenges. These range from self-correcting and self-replicating code through smart forgetting and provable deletion to new frontiers, such as DNA data storage.

Policy and Rule Management - this category focuses on the concrete definition of policies and the distinction between digital preservation policies and rules, guidelines, etc. used in organizations. Some further research challenges are the exploration of policies as boundaries in decision making and negotiations between stakeholders as well as the management of policies that change over time and the association of the versions with different digital objects.

Ethics, Privacy, Security and Trust - the topics of ethics and privacy are closely connected with security and trust and their interplay provides a number of interesting research challenges and questions.

Evaluation and Benchmarking in Digital Preservation The development and improvement of current characterisation and quality assurance techniques is fundamentally hindered by the non-existence of benchmarks. Annotated benchmark data are needed to support the objective comparison of new approaches and quantify the improvements over existing techniques.

Application Domains - as a representative of these applications computer games were chosen. If there is a way to preserve computer games in the long-term it is very likely that all other digital applications will be preservable too. However, this task presents many challenges spreading from hardware through software to copyright and legal issues.

3.4 DP Research Challenges Wiki

Following up on the Dagstuhl seminar, the goal to outline computer science research challenges in DP has inspired the creation of a wiki platform to start a broad discussion on arising challenges involving a widespread global audience. This initiative is currently independent from actual EU-funded projects. This *DP Research WIKI* is publicly available at

<http://socrates.ifs.tuwien.ac.at/wiki/index.php>.

This platform is entirely open and will encourage broad participation. Challenges will be proposed, discussed, refined and published in three phases:

1. A new challenge proposed by a wiki user or submitted via email is introduced through the Incubator for discussion and refinement.

2. After initial discussion, the clarified challenge is moved to one of three sections: Core Computer Science Challenges, Organizational Challenges or Application-oriented Challenges. On these pages, the discussion continues until the specification has reached a level of maturity that merits to freeze the discussion and publish the challenge.
3. Published challenges will be protected to prevent them from changing. They still can be discussed on their talk pages and may be refined later on.

This platform is thus taking forward the discussion on the topics emerging in the Dagstuhl workshop and refining each of them into concrete research challenges with clear motivation, background and research questions. Furthermore, additional topics are being introduced and discussed. The remainder of this section outlines recently added issues in infrastructure-based and content-based research.

3.4.1 Focus of infrastructure-based research

Recent industry analysis suggests that, for the first time, the rate of production of digital data is overtaking the rate of increase in world storage capacity. This fact, combined with the demand for long-term access to this data, will drive two requirements for the digital information life cycle:

Automatic and semi-automatic techniques for storage prioritization (appraisal)

It will become necessary to determine which data should be stored and which should be discarded - and due to the large volumes of data in question (hundreds of Exabytes), automated decision-making is essential. For data that should be persistent, one must determine whether medium-term storage is sufficient or whether the data is sufficiently important to be deposited in a (more expensive) long-term archive. Other data (e.g. IP-packets or video surveillance camera footage) can be identified as transient and demands a different storage policy.

Inexpensive, on-demand storage and processing power

For non-transient data, significant processing power may be required for ingestion into archival storage (if for example media normalization, that is, migration to preservation-friendly formats like PDF/A, is required) or for data management (to support indexing, metadata extraction, and semantic enrichment for efficient retrieval). Both of these requirements demand new applications of established Grid techniques (providing the foundation for distributed storage and computing) as well as economical Cloud Computing approaches (which provide for inexpensive, on-demand access to virtualized computing resources). It is both expensive and inefficient for memory institutions and commercial stakeholders to be concerned with providing the technical infrastructure necessary for long-term archiving. We believe that digital preservation will evolve

to become an external service/infrastructure, offered by specialists to such institutions. Digital curation will remain an area of expertise for librarians and archivists, but they should be able to carry out these functions without at the same time becoming large scale data center administrators and technical experts. We envision global archival services that will be offered by commercial organizations (such as Amazon's EC3/S3 services) as well as governmental organizations (computing centers dedicated to providing infrastructure for the preservation of public goods and cultural heritage). The offered services would include ingest and access, but also the ability to perform preservation actions (feature extraction, indexing, on-demand format migration, etc.) directly within the storage system, circumventing the need for expensive data transfer. This approach also requires the use of standard APIs for utility computing, virtualization, and resource provisioning, such as are under development in the Cloud and Grid communities.

Although European research in the past has assumed that the problem of bit-stream preservation has been solved, this is in fact not the case. Inevitable failures of hardware systems, storage media, and human operators will always endanger the long-term integrity of bits. Therefore, global preservation services must be based on distributed systems with high replication and redundancy. Research in file systems that include massive scalability (for example ZFS) as well as built-in fixity tests and point-to-point data integrity checking is required.

Interoperability between distributed computation and storage networks will remain another research challenge, in particular as nodes in such networks will always include numerous legacy systems at any given time. Finally, research in the area of cyber-security will be required in order to ensure the necessary levels of trust in external preservation services. Policy and regulations regarding IPR, copyright, and privacy must also be taken into consideration.

3.4.2 Focus of content-based research

In terms of content, we note that past projects have concentrated heavily on file-based content (including a significant portion of scientific data that is also file-based). Future research should also consider structured data, in particular databases. Other technical content, targeting specifically the European industry, should include construction and engineering data, pharmaceutical data, and medical records (supporting the development of life-long electronic health records). Also, the preservation of software, in particular open source software archives will become an important policy issue. Technically it is perhaps not so demanding, but as open source renderers are a fundamental aspect for future retrieval of archived resources, this must be considered a high-priority topic.

Furthermore, a new type of digital object will become very significant for accessing archived objects, renderers, and applications - the Virtual Machine Image. Here work must progress towards open standardization of virtual image formats, as well as methods for achieving hardware independence for virtual images.

4 Conclusion

This report gives an overview about research activities on digital preservation within projects co-funded by the European Union in the ICT programme. Eighteen projects, current and past, were analyzed and key objectives were identified. The number of research projects and the large number of deliverables, publications and outcomes made this report a challenging task. The representation of the single projects cannot be complete. The major contributions of the research projects were considered.

This report provides a structured overview about the research projects within project funded by the ICT programme regarding the core objectives, the project consortia and the target communities. Developments and achievements of the European research on digital preservation were identified.

At the outset, the research work was strongly influenced by the library and archive sector. The focus was on the establishment of common problem definitions, terminology, first models and concepts. The primary digital objects were office documents and images from institutional settings. The focus shifted quickly towards more technical aspects of digital preservation and the development of frameworks and tool support for DP. The good networking within the community resulted in continued work across different projects.

The current research moves beyond the preservation of simple documents and data structures to more complex resources such as interactive objects, embedded objects, ontologies and ephemeral data. Just started research projects are addressing the preservation of business process. A general trend in DP is the shift of the forces from the ALM sector towards business and industry sector. One indication of this is the strengthened research on tools, methods and models that work on realistic heterogeneous large collections of complex digital objects.

In the second part of this report research agendas in the field of digital preservation are summarized and challenges for the future are identified. Current research projects have their main focus on issues for the short or medium term. A number of research issues identified in earlier research roadmaps were addressed by research projects in the meantime.

Current initiatives to define future research were addressing more issues in the distant future. The research in DP needs to strengthen the relation and integration with other expertise beyond the core DP disciplines. Examples are technical research on hardware for the specific needs of digital preservation and its application domains or computer science for DP-ready system that integrate long term preservation as a non-foundational requirement from the design perspective. Scalable systems that are able to process and preserve large amounts of data in an efficient and economical way are required.

Distributed and on-demand services will help systems to operate on real life large scale data and provide economically efficient services for customers. Emerging application domains that need to be further addressed by digital preservation research are e-science, e-medicine, e-gov and e-commerce.

5 Appendix

This appendix offers an overview of the projects discussed in the report and provides some key facts about them, as well as a brief description of their objectives.

5.1 APARSEN



Project type:	Network of Excellence
Start date:	01. Jan. 2011
Duration:	48 Months
EU funding:	EUR 6 840 000
Number of partners:	30
URL:	http://www.alliancepermanentaccess.org/current-projects/aparsen

Digital preservation offers the economic and social benefits associated with the long-term preservation of information, knowledge and know-how for re-use by later generations. However, digital preservation has a great problem, namely that preservation support structures are built on projects which are short lived and is fragmented. The unique feature of APARSEN is that it is building on the already established Alliance for Permanent Access (APA), a membership organisation of major European stakeholders in digital data and digital preservation. These stakeholders have come together to create a shared vision and framework for a sustainable digital information infrastructure providing permanent access to digitally encoded information.

To this self-sustaining grouping APARSEN will bring a wide range of other experts in digital preservation including academic and commercial researchers, as well as researchers in other cross-European organisations.

The members of the APA and other members of the consortium already undertake research in digital preservation individually but even here the effort is fragmented despite smaller groupings of these organisations working together in specific EU and national projects. APARSEN will help to combine and integrate these programmes into a shared programme of work, thereby creating the pre-eminent virtual research centre in digital preservation in Europe, if not the World. The APA provides a natural basis for a longer term consolidation of digital preservation research and expertise.

The Joint Programme of Activity will cover:

- technical methods for preservation, access and most importantly re-use of data holdings over the whole lifecycle;
- legal and economic issues including costs and governance issues as well as digital rights;
- outreach within and outside the consortium to help to create a discipline of data curators with appropriate qualifications;

5.2 ARCOMEM



Project type:	Integrated Project
Start date:	01. Jan. 2011
Duration:	36 Months
EU funding:	EUR 6 000 000
Number of partners:	11
URL:	http://www.arcomem.eu

ARCOMEM is about memory institutions like archives, museums, and libraries in the age of the Social Web. Memory institutions are more important now than ever: as we face greater economic and environmental challenges we need our understanding of the past to help us navigate to a sustainable future. This is a core function of democracies, but this function faces stiff new challenges in face of the Social Web, and of the radical changes in information creation, communication and citizen involvement that currently characterise our information society. Social media are becoming more and more pervasive in all areas of life, but this material is both ephemeral and highly contextualised, making it increasingly difficult for a political archivist to decide what to preserve.

The ARCOMEM project team will provide innovative tools for archivists to help exploit the new media and make our organisational memories richer and more relevant. This will be addressed in three ways:

- First, it will be demonstrated how social media can help archivists select material for inclusion, providing content appraisal via the social web.
- Second, it will be shown how social media mining can enrich archives, moving towards structured preservation around semantic categories.
- Third, research will explore social, community and user-based archive creation methods.

The impact of these outcomes will be to a) reduce the risk of losing irreplaceable ephemeral webinformation, b) facilitate cost-efficient and effective archive creation, and c) support the creation of more valuable archives. This has the potential to strengthen our democracies' understanding of the past, in order to better direct our present towards viable and sustainable modes of living, and thus to make a contribution to the future of Europe and beyond.

5.3 BLOGFOREVER



Project type:	STREP
Start date:	01. March. 2011
Duration:	30 Months
EU funding:	EUR 3.16 million
Number of partners:	12
URL:	http://blogforever.eu

BLOGFOREVER will develop robust digital preservation, management and dissemination facilities for weblogs. These facilities will be able to capture the dynamic and continuously evolving nature of weblogs, their network and social structure, and the exchange of concepts and ideas that they foster; pieces of information omitted by current Web Archiving methods and solutions. BLOGFOREVER will lay its foundations on exploring weblog structure and semantics, as well as their interconnections and associations with other web information entities, in order to create a generic weblog data model. This model will be used to define a robust digital preservation policy for weblogs, including interoperability and digital rights management issues. A pilot weblog digital repository will then be developed and validated through a set of case studies. The repository will not only unlock people's and organisations' abilities to access and preserve weblog content but will also enable them to understand its evolving social context over time.

The final output of BLOGFOREVER will be a simple weblog digital archiving solution that any user, user group or institution could use to preserve their weblog(s) and ensure their authenticity, integrity, completeness, usability, and long term accessibility as a valuable cultural, social, and intellectual resource. A multitude of parties will benefit from the project, including libraries and information centres, museums, universities, research institutes, businesses, and bloggers.

The BLOGFOREVER partners will combine and utilise multidisciplinary skills, expertise, and ongoing work in the fields of weblogs analytics, web semantics, social networks, and online preservation. Academic partners will study weblog semantics and the social importance of weblogs; business entities will guarantee the successful take-up and exploitation of the project's outputs. Representatives from bloggers communities will ensure that the results cover their needs. The consortium as a whole is diverse and combines multidisciplinary skills and expertise suitable for the planned research.

5.4 CASPAR



Project type:	Integrated Project
Start date:	01. Apr. 2006
Duration:	42 Months
EU funding:	EUR 8 800 000
Number of partners:	17
URL:	http://www.casparpreserves.eu

CASPAR addressed the growing challenge facing society of a deluge of intrinsically fragile digital information, upon which it is increasingly dependent, by building a pioneering framework to support the end-to-end preservation "life-cycle" for scientific, artistic and cultural information, based on existing and emerging standards. The ambitious challenge to build up a common preservation framework for heterogeneous data and variety of innovative applications was achieved through the following objectives:

- to establish the foundation methodology for covering all preservation aspects. The guiding principle of CASPAR was the application of the OAIS Reference Model
- to research, develop and integrate advanced components to be used in all the preservation activities. These components are the building blocks of the CASPAR Framework
- to create the CASPAR framework: the software platform that enables the building of services and applications that can be adapted to multiple areas and, in particular, to the three testbeds envisaged in the project
- to demonstrate the validity of the CASPAR through heterogeneous testbeds, covering a wide range of disciplines from science to culture to contemporary arts and media, providing a reliable common infrastructure for all.

To achieve this, CASPAR brought together a consortium covering important digital holdings, with the appropriate extensive scientific, cultural and creative expertise, together with commercial partners, and world leaders in the field of information preservation. The consortium is committed to support the further evolution of the framework created.

5.5 DELOS



Project type:	Network of Excellence
Start date:	01. Jan. 2004
Duration:	48 Months
EU funding:	EUR 6 000 000
Number of partners:	57
URL:	http://www.delos.info

The DELOS vision was that digital libraries would become "the universal knowledge repositories and communication conduits for the future, common vehicles by which everyone will access, analyse, evaluate, enhance, and exchange all forms of information. They will be accessible at any time and from anywhere, and will offer a friendly, multi-modal, efficient, and effective interaction and exploration environment".

The main effort of the DELOS Network of Excellence has been towards bridging the gap between this vision and the reality, by furthering research in many critical aspects of digital libraries and by the creation of an active European digital library research community.

The DELOS community, through over 500 scientific papers, has provided significant contributions to many key components of digital libraries, such as advanced and specialised digital library architectures; automatic metadata capturing and extraction from multimedia collections; mechanisms for the integration and automation of appraisal and ingestion of digital material; ontologies for both visual and textual concepts; personalised, context-aware multilingual and multimodal information retrieval, delivery and presentation; user-friendly interfaces; annotation services; testbeds for comparative systems and system component evaluation.

One of the joint activities of the DELOS network was to develop next generation digital library technologies. DelosDLMS, a prototype and demonstrator for future digital libraries, offers various services and specialised functionalities on top of a reliable and scalable middleware infrastructure.

Another important research challenge has been the interoperability of the various content holders, i.e. the ability to store and retrieve information across collections in diverse media and languages. The DELOS contribution to this domain was the development of a digital library reference model.

5.6 DPE



Project type:	Coordination Action
Start date:	01. Apr. 2006
Duration:	36 Months
EU funding:	EUR 1 451 000
Number of partners:	11
URL:	http://www.digitalpreservationeurope.eu

The Coordination Action DigitalPreservationEurope was launched in order to improve cooperation and consistency in current activities to secure effective preservation of digital materials, and to help both citizens and specialist professionals recognise the central role that digital preservation plays in their lives and work. To this end, the project has facilitated pooling of the complementary expertise that exists across the academic research, cultural institutions, public administrations and industry sectors in Europe.

Project results:

- DPE released the DRAMBORA toolkit for repository auditing and PLATTER ('Planning Tool for Trusted Electronic Repositories') which provides a basis for a digital repository to plan the development of its goals, objectives and performance targets.
- Within the work package 'Coordination of EU Repository Activities', DPE has designed a Registry of Digital Repositories in order to monitor and assess information on preservation policies and practices of organisations. Further, DPE has launched a service to issue unique identifiers for digital objects and produced a guidance document for repository planners seeking to achieve trusted status, consistent with internationally accepted standards for repository management.
- Among the DPE publications are a state of the art review on international competence centres for digital curation and preservation activities and expertise; a 'Market and Technology Trends Analysis' on needs and plans of main stakeholders and technological solutions available for digital preservation; and a 'Digital Preservation Research Roadmap' identifying core domains for preservation research.
- DPE also has established training programmes on digital preservation, run several training courses and created online training materials.

5.7 ENSURE



Project type:	Integrated Project
Start date:	01. Feb. 2011
Duration:	36 Months
EU funding:	EUR 7.85
Number of partners:	13
URL:	http://ensure-fp7-plone.fe.up.pt

Ensuring long term usability for the spiralling amounts of data produced or controlled by organizations with commercial interests is quickly becoming a major problem. Drawing on motivation from use cases in health care, finance and clinical trials, ENSURE will significantly extend the state of the art in digital preservation which to-date has focused on relatively homogeneous cultural heritage data. Our use cases bring up a large number of issues which have yet to be fully addressed:

1. safely leveraging scalable pay-as-you-go infrastructure such as clouds
2. having businesses understand the economic implications of preservation,
3. conforming to regulatory, contractual and legal requirements as part of a whole workflow
4. managing long term integrity and authenticity significant intellectual property or highly personal data and
5. using off-the-shelf IT technologies for preservation to support different types of digital resources.

Building on prior work, ENSURE will address these issues with innovative approaches and tools: Cost and Value Evaluate the cost and benefit of different quality solutions. Preservation Lifecycle Management Build on industry standard lifecycle management approaches to manage the preservation lifecycle, ensuring regulatory compliance, allowing changes in the preservation approach to reflect environmental changes, addressing evolution of ontologies and managing the quality of the digital objects over time.

5.8 ERPANET



Project type:	Preparatory, accompanying and support measures
Start date:	01. Nov. 2001
Duration:	36 Months
EU funding:	EUR 899 000
Number of partners:	3
URL:	http://www.erpanet.org

This network established an expandable and self-sustaining European initiative, which aims to serve as a virtual clearinghouse and knowledge-base in the area of preservation of cultural heritage and scientific digital objects. The dominant feature of ERPANET is the exchange of knowledge on state-of-the-art developments in digital preservation and the transfer of expertise among individuals and institutions. More specifically, ERPANET delivers a range of services (e.g. content creation, advisory service, training and thematic workshops and fora), both to information creation and user community. It makes accessible tools, knowledge, and experience.

ERPANET did not directly carry out new research to develop such tools, but it created a coherent platform for proactive co-operation, collaboration, exchange and dissemination of research results and experience in the preservation of digital objects. The project consortium brought together research institutions, memory organisations, ICT industry, entertainment and creative (e.g. broadcasting) industries with the goal to provide an effective, multidisciplinary, knowledge and resource-sharing infrastructure.

5.9 KEEP



Project type:	Integrated Project
Start date:	01. Feb. 2009
Duration:	36 Months
EU funding:	EUR 3 150 000
Number of partners:	8
URL:	http://www.keep-project.eu

KEEP (Keeping Emulation Environments Portable) is developing emulation services (KEEP Emulation Services) to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames etc.

The overall aim of the project is to facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects. KEEP is also considering legal issues concerning the implementation of emulation-based systems and will propose solutions which comply with European and national copyright laws.

KEEP addresses the problems of transferring digital objects stored on outdated computer media such as floppy discs onto current storage devices. This involves the specification of file formats and the production of transfer tools exploited within a framework, and taking into account possible legal and technical issues. KEEP addresses all aspects ranging from safeguarding the original bits from the carrier to offering online services to end-users via a highly portable Emulation Framework running on any possible device. In addition to producing a software package, the project is delivering understanding knowledge about how to integrate emulation-based solutions with an operational electronic deposit system. Existing metadata models are being researched and guidelines developing for mapping digital objects to emulated manifestations. KEEP is seeking ways to integrate its work with the outputs of other digital preservation projects and software (for example Planets and Pronom). Overall, KEEP will contribute to the next generation of permanent access strategies based on emulation.

Although primarily aimed at those involved in Cultural Heritage, such as memory institutions and games museums, the KEEP Emulation Services can also serve the needs of a wide range of organisations and individuals because of its universal approach.

5.10 LiWA



Project type:	STREP
Start date:	01. Feb. 2008
Duration:	36 Months
EU funding:	EUR 2 682 000
Number of partners:	8
URL:	http://www.liwa-project.eu

The interest in Web content preservation is strongly growing, not only in traditional library and archival organisations, but also in sectors such as industry and services. But the typical characteristics of Web content - variety of formats, high dynamics, volatility, interactivity and context-dependency - make adequate Web archiving a particular challenge. With the LiWA project, Web archiving has been established as a new topic for scientific research and development within the digital preservation domain.

At the centre of the project was the concept of 'Living Web Archives', as opposed to the current practice of producing periodic snapshots of pages. 'Living' here refers to: * long term interpretability as the archive evolves and adapts over time, * improved archive fidelity and authenticity by filtering out irrelevant information, * captured content from a wide variety of sources.

To enhance archive fidelity and authenticity, LiWA has developed and tested new methods based on content interpretation and intelligent pattern detection of traps and Web spam. This allows reducing the amount of fake content and helping prioritise crawls by automatically detecting content of value. To improve the integrity and temporal, structural and semantic coherence of Web archives, some work was dedicated to temporal Web archive construction. This serves the objective to significantly improve content positioning in time and (topic) space and will lay the foundations for fast and effective access to evolving Web content.

To facilitate archive interpretability, LiWA applied methods for semantic and terminology extraction, able to detect and handle evolving semantics, interpretations of domain concepts and terminology. This is a contribution to the task of preserving the usefulness, quality, and accessibility of Web archives over time. For validating the LiWA approach, two demonstrator applications have been built on top of the LiWA services. The applications focus on the social Web and on the special challenge of archiving audio-visual content.

5.11 PARSE.Insight*



Project type: FP7 INFRASTRUCTURES
Start date: 01. March 2008
Duration: 24 Months
EU funding: EUR 1 250 000
Number of partners: 10
URL: <http://www.parse-insight.eu>

PARSE.Insight aimed to highlight the longevity and vulnerability of digital research data and concentrates on the parts of the e-Science infrastructure needed to support persistence and understandability of the digital assets of EU research.

PARSE.Insight was concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research. The problem is how to safeguard this valuable digital material over time, to ensure that it is accessible, usable and understandable in future. The rapid pace of change in information technology threatens media, file formats and software with obsolescence, and changing concepts and terminology also mean that, even if data can be read, it might not be correctly interpreted by future generations.

Many initiatives are already under way in this area, and the aim of the PARSE.Insight project was to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The project conducted surveys and in-depth case studies of different scientific disciplines and stakeholders and bases its results on these findings, as well as knowledge of ongoing developments.

* co-funded under FP7-INFRASTRUCTURES

5.12 PLANETS



Project type:	Integrated Project
Start date:	01. Jun. 2006
Duration:	48 Months
EU funding:	EUR 8 600 000
Number of partners:	16
URL:	http://www.planets-project.eu

The primary goal for PLANETS was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets. The project delivered an integrated production environment for the management of digital information preservation, with a special focus on the needs of libraries and archives.

The PLANETS environment supports a number of key preservation functions through:

- the preservation planning tool PLATO and services that empower organisations to define, evaluate, and execute high-quality and cost-effective preservation plans
- methodologies, tools and services for the characterisation of digital objects that can automatically analyse digital objects and establish significant properties
- innovative solutions for performing preservation actions and to ensure rendering of the objects and keeping their properties available. In this context, work has been done on the archiving of relational databases, emulation and remote access to emulation services.

Integration and automation can be seen as the two prominent features of the Planets environment. The PLANETS Interoperability Framework integrates the deliverables into a downloadable 'click and install' software package. Within this package, there are role-based routines for administrators, preservation experts and business users, enabling organisations to improve decision-making about long term preservation, ensure long-term access to their valued digital content and control the costs of preservation actions.

5.13 PrestoPRIME



Project type:	Integrated Project
Start date:	01. Jan. 2009
Duration:	42 Months
EU funding:	EUR 8 000 000
Number of partners:	14
URL:	http://www.prestoprime.org

PrestoPRIME researches and develops practical solutions for the long-term preservation of digital media objects, programmes and collections, and will find ways to increase access to them. The project will deliver a preservation framework, complete with risk management and content quality and corruption control measures, capable of supporting audiovisual signal migration and multivalent preservation methods using federated services for distributing and storing content. A metadata conversion and deployment toolkit will be generated, supporting a novel and efficient process for metadata vocabulary alignment, annotation and services for user-generated content metadata. A rights management system and audiovisual fingerprint registry will make it possible to track and manage content at all stages of its lifecycle, in all contexts of use.

The project activities are guided by four objectives: (1) to research and develop means of ensuring the permanence of digital audiovisual content in archives, libraries, museums and other collections; (2) to research and develop means of ensuring the long-term future access to audiovisual content in dynamically changing contexts; (3) to integrate, evaluate and demonstrate tools and processes for audiovisual digital permanence and access; (4) to establish a European networked Competence Centre to gather the knowledge created through the research collaboration and share it with the stakeholder community.

An important achievement of the project team so far is the launch of the networked competence centre, branded "PrestoCentre" in March 2011. PrestoCentre was created to, enhance collaboration between audiovisual content holders in Europe; facilitate coordinated action in the areas of digitisation, digital preservation of and long term access to audiovisual archival content; and serve an international community of stakeholders in audiovisual digitisation and digital preservation through online and offline services, publications and training.

5.14 PROTAGE



Project type:	STREP
Start date:	01. Nov. 2007
Duration:	36 Months
EU funding:	EUR 2 021 000
Number of partners:	7
URL:	http://www.protage.eu

PROTAGE addressed the challenges related to the preservation of digital resources of increasing volume and heterogeneity. The solution proposed was to develop tools allowing for more efficiency and self-reliance of preservation processes.

For this purpose, PROTAGE researchers explored the value of a promising technology - software agents - for the automation of digital preservation processes. Based on the latest research on digital preservation strategies and on autonomous systems, the project has built and validated flexible and extensible software agents for long-term digital preservation and access that can cooperate with and be integrated in existing and new preservation systems to support various aspects of the digital preservation workflow such as the submission / ingestion of digital material, monitoring of preservation systems and transfer between repositories. Tools developed by the PROTAGE project will:

- enable content producers to create and publish in a preservation-compatible manner,
- provide digital repositories with means of further automating the preservation processes,
- facilitate seamless interoperation between content providers, libraries and archives, and end-users throughout Europe.

Targeted end users are curators and digital content creators, including individuals managing their own digital collections. PROTAGE will use archive and library materials from the project partners for system and user tests and external stakeholders in further validation. The Swedish Centre of Competence for Long-term Preservation will ensure availability of results to a wider community of memory institutions. The industrial partners will use the results to develop commercial solutions.

5.15 SCAPE



Project type:	Integrated Project
Start date:	01. Feb. 2011
Duration:	42 Months
EU funding:	EUR 8 600 000
Number of partners:	14
URL:	http://www.scape-project.eu

The SCAPE project will enhance the state of the art of digital preservation in three ways: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system.

These concrete project results will be validated within three large-scale Testbeds from diverse application areas: Digital Repositories from the library community, Web Content from the web archiving community, and Research Data Sets from the scientific community. Each Testbed has been selected because it highlights unique challenges. SCAPE will develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects. These services will be able to

- Identify requirements for preserving all or parts of a repository through characterisation and trend analysis
- Define responses to those needs using formal descriptions of preservation policies and preservation plans
- Allow a high degree of automation, virtualization of tools, and scalable processing
- Monitor the quality of preservation processes.

The SCAPE consortium brings together experts from memory institutions, data centres, research labs, universities, and industrial firms in order to research and develop scalable preservation systems that can be practically deployed within the project lifetime. SCAPE is dedicated toward producing open source software solutions available to the entire digital preservation community.

5.16 SHAMAN



Project type: Integrated Project
Start date: 01. Dec. 2007
Duration: 48 Months
EU funding: EUR 8 398 000
Number of partners: 18
URL: <http://shaman-ip.eu>

This project will develop and test a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives. The aim of SHAMAN is to develop the framework for the next generation of long term (more than one century) digital preservation systems and tools. This includes the definition of a theory of preservation that integrates the analysis, ingestion, management, access to and reuse of information objects across distributed repositories. The data preservation capabilities offered will secure the authenticity and integrity of data objects over time. The development work will be structured around four core components. Their objectives can be described as follows:

- to establish an open distributed resource management infrastructure framework enabling grid-based resource integration, reflecting, refining and extending the OAI model and taking advantage of the latest state of the art in virtualisation and distribution technologies from the fields of GRID computing, Federated Digital Libraries, and Persistent Archives;
- to develop and integrate technologies to support contextual and multi-valent archival and preservation processes which are adapted and significantly extended from the fields of content and document Management and Information Systems;
- to develop and integrate technologies to support semantic constraint-based collection management to target one of the key challenges in automating one class of digital preservation core functions;
- to support the managing of future requirements by securing interoperability with future environments and maintaining essential properties of the preserved content.

5.17 TIMBUS



Project type:	Integrated Project
Start date:	01. April 2011
Duration:	36 Months
EU funding:	EUR 7.78 million
Number of partners:	10
URL:	http://timbusproject.net

The digital preservation problem is well-understood for query-centric information scenarios but has been less explored for scenarios where the important digital information to be preserved is the execution context within which data is processed, analysed, transformed and rendered. Furthermore, preservation is often considered as a set of activities carried out in the isolation of a single domain, without considering the dependencies on third-party services, information and capabilities that will be necessary to validate digital information in a future usage context.

A primary motivation for TIMBUS is the declining popularity of centralized in-house business processes maintained and owned by single entities. The presence of Software as a Service (SaaS) and Internet of Services (IoS) means business processes are increasingly supported by service oriented systems where numerous services provided by different providers, located in different geographical locations are composed to form value added service compositions and service systems which will continue changing and evolving. Besides the advantages of SaaS and IoS there is the danger of services and service providers disappearing (for various reasons) leaving partially complete business processes.

TIMBUS will endeavour to enlarge the understanding of DP to include the set of activities, processes and tools that ensure continued access to services and software necessary to produce the context within which information can be accessed, properly rendered, validated and transformed into context based knowledge. One of the fundamental requirements is to preserve the functional and non-functional specifications of services and software, along with their dependencies. This is more challenging than the plain preservation of data as elements including, but not limited to, the versioning, licensing, cryptographic schemes, known data formats, host-system environments, architectures and hardware requirements of software continue to change over time. This enlarged understanding brings DP clearly into the domain of Business Continuity Management (BCM).

5.18 WF4EVER



Project type:	STREP
Start date:	01. Dec. 2010
Duration:	36 Months
EU funding:	EUR 3.86 million
Number of partners:	6
URL:	http://www.wf4ever-project.org

Wf4Ever aims at providing the methods and tools required to ensure the long-term preservation of scientific workflows in order to support the scientific discovery process and the development of new scientific assets. Wf4Ever will develop new models, techniques and tools for the preservation of scientific workflows, including the novel definition of a Research Object, which packages workflow descriptions, the provenance of their executions, and links to all the related resources upon which they depend. Such models will also include models for repeatability and reproducibility, and models for workflow abstraction, to facilitate workflow classification and indexing, comparison, and similarity detection between pairs of existing workflows in the library. Wf4Ever will also develop strategies for sharing and reusing workflows or workflow fragments and patterns, including mechanisms for personalised workflow recommendation based on workflow descriptions, users collective behaviour, and social information.

Finally, Wf4Ever will propose methods and tools to proactively preserve and inspect workflow integrity and authenticity through the evaluation of workflow information quality, based on the provenance of workflows and their research objects and described in new vocabularies for the representation of the provenance of research objects in digital preservation systems. Wf4Ever will develop a software architecture and reference implementation for the preservation of scientific workflows, which will extend one of the most widely deployed scientific workflow sharing infrastructures (myExperiment) with preservation capabilities that consider the complexity of scientific workflows and their related objects. This software system will thus leverage the advances done on workflow lifecycle management, collaboration and sharing support, and integrity and authenticity maintenance. Wf4Ever will be evaluated in two workflow-intensive use cases in the domains of Astronomy and Genomics.

References

- [1] Daniel Gelaw Alemneh and Samantha Kelly Hastings. Exploration of adoption of preservation metadata in cultural heritage institutions: case of premis. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 50:1–50:8, Silver Springs, MD, USA, 2010. American Society for Information Science.
- [2] Jose Barateiro, Goncalo Antunes, and Jose Borbinha. Addressing digital preservation: Proposals for new perspectives. In *Proceedings of First International Workshop on Innovation in Digital Preservation (InDP)*, 2009.
- [3] CASPAR consortium. D1201: Conceptual Model. Public deliverable, May 2007.
- [4] Jean-Pierre Chanod, Milena Dobрева, Andreas Rauber, Seamus Ross, and Vittore Casarosa. 10291 Report – Automation in Digital Preservation. In *Automation in Digital Preservation*, number 10291 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.
- [5] Angela Dappert and Markus Enders. Digital preservation metadata standards. *Information Standards Quarterly*, 22(2), 2010.
- [6] Angela Dappert and Adam Farquhar. Implementing metadata that guides digital preservation services. In *Proceedings of the 6th iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, pages 50 – 58. California Digital Library, 2009.
- [7] Angela Dappert and Adam Farquhar. Significance is in the eye of the stakeholder. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL 09)*, volume 5714 of *Lecture Notes in Computer Science*, pages 297–308. Springer Berlin / Heidelberg, 2009.
- [8] Michael Day. Metadata for digital preservation: A review of recent developments. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 161–172, London, UK, 2001. Springer Berlin / Heidelberg.
- [9] DigitalPreservationEurope consortium. D7.2 -Research Roadmap. Public deliverable, June 2006.
- [10] Susan S. Lazinger and Helen R. Tibbo. *Digital Preservation and Metadata: History, Theory, Practice*. Libraries Unlimited, Inc., Englewood, CO, USA, 2001.
- [11] PARSE.Insight consortium. Deliverable D2.2 - Science Data Infrastructure Roadmap. Public deliverable, June 2010.

- [12] PrestoPRIME consortium. Deliverable 5.2.1 - Definition and Design of a PrestoPRIME Reference Architecture for the Integration Framework. Public deliverable, June 2010.
- [13] Andreas Rauber, Stephan Strodl, Carl Rauch, Hans Hofman, Giuseppe Amato, Max Kaiser, and Heike Neuroth. DELOS DPC Testbed. DELOS Research Activities 2005, July 2005.
- [14] SHAMAN Consortium. D2.3 Specification of the SHAMAN Reference Architecture. Public deliverable, May 2009.